Atty. Docket No. __16528X-025010__

"Express Mail" Label No. __EM197111645US__

Date of Deposit __5/14/97__

I hereby certify that this is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Asst. Commissioner for Patents, Washington, D. C. 20231

By __/Ka M. Vallkan C__

PATENT APPLICATION
ASST. COMMISSIONER FOR PATENTS
Washington, D. C. 20231

Sir:

Transmitted herewith for filing is the
   [X] patent application of
   [ ] design patent application of
   [ ] continuation-in-part patent application of

Inventor(s): Mark Chee, Anthony Berno and Robert Yang

For: POLYMORPHISMS IN HUMAN MITOCHONDRIAL DNA

[X] This application claims priority from each of the following Application Nos./filing dates:
   __60/017,203 / May 16, 1996__ and __60/024,206 / August 20, 1996.__

[ ] Please amend this application by adding the following before the first sentence: --This application claims the benefit of U.S. Provisional Application No. 60/_____ , filed _____ , the disclosure of which is incorporated by reference.--

Enclosed are:

[X] Application with 29 pages of Specification, 2 pages of claims and 1 page of abstract.
[ ] _3_ sheets informal drawing(s).
[X] A unsigned Declaration & Power of Attorney.
[ ] A [ ] signed [ ] unsigned Declaration.
[ ] A Power of Attorney.
[ ] A verified statement to establish small entity status under 37 CFR 1.9 and 37 CFR 1.27 [ ] is enclosed [ ] was filed in the earliest of the above-identified patent application(s).
[ ] A certified copy of a _____ application.
[ ] Information Disclosure Statement under 37 CFR 1.97.
[ ] A petition to extend time to respond in the parent application of this continuation-in-part application.
[ ] Return postcard

In view of the Unsigned Declaration as filed with this application and pursuant to 37 CFR §1.53(d), Applicant requests deferral of the filing fee until submission of the Missing Parts of Application.

DO **NOT** CHARGE THE FILING FEE AT THIS TIME.

Telephone:
(415) 576-0200

m:\home\jol\work\16528\250-1\app.tm

Joe Liebeschuetz
Reg. No.: 37,505
Attorneys for Applicant

PATENT APPLICATION

POLYMORPHISMS IN HUMAN MITOCHONDRIAL DNA

Inventor: Mark Chee
                 3199 Waverly Street
                 Palo Alto, California   94306
                 Citizen of Australia

                 Anthony Berno
                 570 South 12th Street
                 San Jose, California   95112
                 Citizen of Canada

                 Robert Yang
                 317 Winged Terrace
                 San Ramon, California   94583
                 Citizen of the United States

         Assignee: Affymetrix, Inc.

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California   94111-3834
(415) 576-0200

CROSS-REFERENCE TO RELATED APPLICATIONS
5   The present application derives priority from USSN
60/017,203 filed May 16, 1996 and USSN 60/024,206, filed
August 20, 1996, which are incorporated by reference in their
entirety for all purposes.

10   STATEMENT OF GOVERNMENT INTEREST
       This application was funded in part by SPIR grant no.
R43HGO1481-01 and NIH grant no. 5r01HG00813. The US
Government may have certain rights in this invention.

15   BACKGROUND
       Human mitochondrial DNA (mtDNA) is information-rich,
encoding some 22 tRNAs, a 12S and a 16S rRNA, and 13
polypeptides involved in oxidative phosphorylation. No
introns have been detected. RNAs are processed by cleavage at
20   tRNA sequences, and polyadenylated posttranscriptionally. In
some transcripts, polyadenylation also creates the stop codon,
illustrating the parsimony of coding.
       A complete prototypical sequence of the human
mitochondrial genome has been published. See Anderson et al.,
25   *Nature* 290, 457-465 (1981). The reported sequence in 16,569
base pairs long. There is strand asymmetry in the base
compositions, with one strand (Heavy) being relatively G rich,
and the other strand (Light) being C rich. The L strand is -
30.9% A, 31.2% C, 13.1% G, and 24.7% T. The sequence of the
30   L-strand is numbered arbitrarily from the MboI-5/7 boundary in
the D-loop region.
       A human cell may have several hundred or more
mitochondria, each with more than one copy of mtDNA. Hence
each cell actually contains a population of mtDNA molecules.
35   In many individuals, the mtDNA sequence is essentially clonal.
However, some individuals carry more than one mtDNA sequence,
a condition known as heteroplasmy. The degree of
heteroplasmy can vary from tissue to tissue. Also, the rate

of replication of mtDNAs can differ and together with random
segregation during cell division, can lead to changes in
heteroplasmy over time.

5   Mitochondrial DNA is maternally inherited, and has a
mutation rate estimated to be tenfold higher than single copy
nuclear DNA (Brown et al., *Proc. Natl. Acad. Sci. USA* 76,
1967-1971 (1979)). Over 80% of substitutions are transitions
(*i.e.*, pyrimidine-pyrimidine or purine-purine).

       The determination of a complete human mitochondrial DNA
10   sequence over 15 years ago has had a tremendous influence on
studies of human origins and evolution, and the role of
mutations in degenerative diseases. Cann et al. *Nature* **325**,
31-36 (1987); Zeviani, et al., *Am. J. Hum. Genet.* **47**, 904-914
(1990); Wallace, *Annu. Rev. Biochem.* **61**, 1175-1212 (1992);
15   Horai et al., *Proc. Natl. Acad. Sci. USA* **92**, 532-536 (1995);
Hutchin & Cortopassi, *Proc. Natl. Acad. Sci. USA* **92**, 6892-6895
(1995). Because of the cost and difficulty of conventional
sequence analysis, most sequencing studies have focused only
on two small hypervariable regions totalling ˉ600 bp.
20   Greenberg et al. *Gene* **21**, 33-49 (1983); Aquardo & Greenberg,
Genetics **103**, 287-312 (1983).

       The present application describes sequencing of the
complete mitochondrial genome of several individuals and
identifies a large set of polymorphisms.

25

SUMMARY OF THE INVENTION
       In one aspect, the invention provides nucleic acids
comprising segments of human mitochondrial DNA or RNA of
between 10 and 100 bases. The segments include any one of the
30   182 polymorphic sites shown in Table 1. Also included are
nucleic acids comprising complements of these segments. In
some segments, the polymorphic site is occupied by the base
described by Anderson et al., supra. In other segments, the
site is occupied by a different base, particularly one of the
35   alternative forms shown in Table 1, column 2, or 4-11.

       The invention further provides allele-specific
oligonucleotides for analysis of the polymorphic sites shown
in Table 1. The allele specific oligonucleotides hybridize to

a segment of human mitochondrial nucleic acid or its
complement including a polymorphic site shown in Table 1,
column 1. Such oligonucleotides can be used as probes or
primers.

5   The invention further provides isolated nucleic acids
comprising a segment of the human mitochondrial sequence
described by Anderson et al., *Nature* 290, 457-465 (1981), or
the complement thereof, including a polymorphic site shown in
Table 1, column 1. In these nucleic acids, the polymorphic
10   site within the segment is occupied by a base other than the
base shown in Table 1, column 3 ("asn base").

The invention further provides methods of analyzing a
nucleic acid. Such methods entail obtaining a mitochondrial
nucleic acid from an individual, and determining a base
15   occupying any one of the polymorphic sites shown in Table 1.

BRIEF DESCRIPTION OF THE FIGURES
Fig. 1. (A) Design of a 4L tiled array. Each position in the
target sequence (upper case) is queried by a set of 4 probes
on the chip (lower case), identical except at a single
20   position, termed the substitution position, which is either A,
C, G, or T (blue indicates complementarity, red a mismatch).
Two sets of probes are shown, querying adjacent positions in
the target. (B) Effect of a change in the target sequence.
25   The probes are the same as in panel A, but the target now
contains a single base substitution (C, shown in green). The
probe set querying the changed base still has a perfect match
(the G probe). However, probes in adjacent sets that overlap
the altered target position now have either one or two
30   mismatches (red), instead of zero or one, since they were
designed to match the target shown in panel A. (C)
Hybridization to a 4L tiled array and detection of a base
change in the target. The array shown was designed to the mt1
sequence. (Upper panel) hybridization to mt1. The
35   substitution used in each row of probes is indicated to the
left of the image. The target sequence can be read 5' to 3'
from left to right as the complement of the substitution base
with the brightest signal. With hybridization to mt2 (lower

panel), which differs from mt1 in this region by a T -> C
transition, the G probe at position 16,493 is now a perfect
match, with the other three probes having single base
mismatches (A 5, C 3, G 37, T 4 counts). However, at flanking
positions, the probes have either single or double base
mismatches, since the mt2 transition now occurs away from the
query position.

Fig. 2. Detection of base differences in a 2.5 kb region by
comparison of scaled P^0 hybridization intensity patterns
between a sample (green) and a reference (red). **(A)**
Comparison of ief007 to mt1. In the region shown, there is a
single base difference between the two sequences, located at
position 16,223 (mt1 C: ief007 T). This results in a
"footprint" spanning ~ 20 positions, 11 to the left and 8 to
the right of position 16,223, in which the ief007 P^0
intensities are decreased by a factor of more than 10 on
average relative to the mt1 intensities. The predicted
footprint location is indicated by the gray bar, and the
location of the polymorphism is shown by a vertical black line
within the bar. **(B)** Comparison of ha001 to mt1. The ha001
target has 4 polymorphisms relative to mt1. The P^0 intensity
pattern clearly shows two regions of difference between the
targets. Each region contains at least 2 differences, because
in both cases the footprints are longer than 20 positions, and
therefore are too extensive to be explained by a single base
difference. The effect of competition can be seen by
comparing the mt1 intensities in the ief007 and ha001
experiments: the relative intensities of mt1 are greater in
panel B where ha001 contains P^0 mismatches but ief007 does
not. **(C)** The ha004 sample has multiple differences to mt1,
resulting in a complex pattern extending over most of the
region shown. Thus, differences are clearly detected. Because
hybridization intensities are extremely sequence dependent,
each of the mitochondrial sequences can also be identified
simply by its hybridization pattern.

Fig. 3. Human mitochondrial genome on a chip. (**A**) An image of
the array hybridized to a 16.6 kb L strand mitochondrial
target RNA. The 16,569 bp map of the genome is shown and the
H strand origin of replication (O_H), located in the control
region, is indicated. (**B**) A portion of the hybridization
pattern, magnified. In each column there are 5 probes: A, C,
G, T and D, from top to bottom. The D probe has a single base
deletion instead of a substitution, and hence is 24 instead of
25 bases in length. The scale is indicated by the bar beneath
the image. Although there is considerable intensity variation,
most of the array can be read directly. The image was
collected at a resolution of ˜ 100 pixels per probe cell. (**C**)
The ability of the array to detect and read single base
differences in a 16.6 kb sample is illustrated. Two different
target sequences were hybridized in parallel to different
chips. The hybridization patterns are compared for four
different positions in the sequence. Only the $p^{25,13}$ probes are
shown. The top panel of each pair shows the hybridization of
the mt3 target, which matches the chip P^0 sequence at these
positions. The lower panel shows the pattern generated by a
sample from a patient with Leber's hereditary optic neuropathy
(LHON). Three known pathogenic mutations, LHON3460, LHON4216,
and LHON13708, are clearly detected. For comparison, the
fourth panel in the set shows a region around position 11,778
that is identical in both samples.

DEFINITIONS

Nucleic acids, including oligonucleotides, can be
DNA or RNA, and single- or double-stranded. Oligonucleotides
can be naturally occurring or synthetic, but are typically
prepared by synthetic means. Preferred oligonucleotides of
the invention include segments of DNA, or their complements
including any one of the polymorphic sites shown in Table 1.
The segments are usually between 5 and 100 bases, and often
between 5-10, 5-20, 10-20, 10-50, 20-50 or 20-100 bases. The
polymorphic site can occur within any position of the segment.

The segments can be from any of the allelic forms of DNA shown in Table 1.

Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al., *Science* 254, 1497-1500 (1991).

The term primer refers to a single-stranded oligonucleotide capable of acting as a point of initiation of template-directed DNA synthesis under appropriate conditions (*i.e.*, in the presence of four different nucleoside triphosphates and an agent for polymerization, such as, DNA or RNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. The appropriate length of a primer depends on the intended use of the primer but typically ranges from 15 to 30 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not reflect the exact sequence of the template but must be sufficiently complementary to hybridize with a template. The term primer site refers to the area of the target DNA to which a primer hybridizes. The term primer pair means a set of primers including a 5' upstream primer that hybridizes with the 5' end of the DNA sequence to be amplified and a 3', downstream primer that hybridizes with the complement of the 3' end of the sequence to be amplified.

Linkage describes the tendency of genes, alleles, loci or genetic markers to be inherited together as a result of their location on the same chromosome, and can be measured by percent recombination between the two genes, alleles, loci or genetic markers.

Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length

polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The

5   first identified allelic form is arbitrarily designated as a the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous

10 or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms.

A single nucleotide polymorphism occurs at a polymorphic site occupied by a single nucleotide, which is the site of variation between allelic sequences. The site is

15 usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations).

A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the

20 polymorphic site. A transition is the replacement of one purine by another purine or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine or vice versa. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion

25 of a nucleotide relative to a reference allele.

Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM

30 EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations.

An isolated nucleic acid means an object species invention that is the predominant species present (*i.e.*, on a molar basis it is more abundant than any other individual

35 species in the composition). Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90 percent (on a molar basis) of all macromolecular species present. Most preferably, the object species is purified to essential

homogeneity (contaminant species cannot be detected in the composition by conventional detection methods).

DETAILED DESCRIPTION

I.   Novel Polymorphisms of the Invention

The present application provides the location of 505 polymorphisms at 182 sites in human mitochondrial DNA (see attached Table 1). The first column of the table identifies the position of polymorphisms according to a convention whereby nucleotides in the mitochondrial genomes of different individuals are assigned the same number as the corresponding nucleotide in the sequence of Anderson et al., *supra* when the two sequences are maximally aligned. For example, reading from the top of the table, polymorphisms occur at positions 63, 72, 92, 181, 194 and so on. The column headed rbase indicates the base occupying the position in a mitochondrial reference sequence whose sequence was determined by ABI sequencing. The column headed asn-base indicates the base occupying the position in the published sequence of Anderson et al. (1981), *supra*. The remaining columns in the table indicate the base present at the polymorphic position in each of 10 African individuals (designated from left to right HA001, HA002, HA004 and so forth). Upper case letters indicate that the base occupying a polymorphic position has been established with substantial certainty in an individual. Lower case letters indicate that although the existence of a polymorphic position is certain, there is a possibility of experimental error in determining the base indicated as occupying that position in an individual. The letter "n" indicates that the base occupying a position in an individual was ambiguous. Some ambiguities may be due to heteroplasmy.

The polymorphic sites listed in the table are widely distributed throughout the mitochondrial genome, and occur in many of the known genes within the genome (i.e., 22 tRNAs, a 12S and a 16S rRNA, and 13 polypeptides involved in oxidative phosphorylation). The polymorphic sites can be classified in sets according to which gene the polymorphic site occurs.

II.   Analysis of Polymorphisms

    A.   Preparation of Samples

       Polymorphisms are detected in a target nucleic acid from an individual being analyzed. For assay of mitochondrial DNA, virtually any biological sample is suitable. For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair.

       Amplification of nucleic acids from target samples is sometimes desirable and can be accomplished by e.g., PCR. *See generally PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202 (each of which is incorporated by reference for all purposes).

       Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA,* 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

       Amplification is not always necessary in view of the multiple copies of mitochondrial DNA in a single cell. Analysis without amplification is preferred for analysis of relative levels of various mitochondrial DNA subpopulations, such as mitochondrial DNAs bearing deletions, without affecting their relative levels.

B.   Detection of Polymorphisms in Target DNA

There are two distinct types of analysis depending whether a polymorphism in question has already been characterized. The first type of analysis is sometimes referred to as de novo characterization. This analysis compares target sequences in different individuals to identify points of variation, i.e., polymorphic sites. By analyzing a groups of individuals representing the greatest ethnic diversity among humans and greatest breed and species variety in plants and animals, patterns characteristic of the most common alleles/haplotypes of the locus can be identified, and the frequencies of such populations in the population determined. Additional allelic frequencies can be determined for subpopulations characterized by criteria such as geography, race, or gender. The de novo identification of the polymorphisms of the invention is described in the Examples section. The second type of analysis is determining which form(s) of a characterized polymorphism are present in individuals under test. There are a variety of suitable procedures, which are discussed in turn.

1.   Allele-Specific Probes

The design and use of allele-specific probes for analyzing polymorphisms is described by e.g., Saiki et al., *Nature* 324, 163-166 (1986); Dattagupta, EP 235,726, Saiki, WO 89/11548. Allele-specific probes can be designed that hybridize to a segment of target DNA from one individual but do not hybridize to the corresponding segment from another individual due to the presence of different polymorphic forms in the respective segments from the two individuals. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Some probes are designed to hybridize to a segment of target DNA such that the polymorphic site aligns with a central position (e.g., in a 15 mer at the 7 position; in a 16 mer, at either the 8 or 9 position) of the probe. This design

of probe achieves good discrimination in hybridization between different allelic forms.

Allele-specific probes are often used in pairs, one member of a pair showing a perfect match to a reference form of a target sequence and the other member showing a perfect match to a variant form. Several pairs of probes can then be immobilized on the same support for simultaneous analysis of multiple polymorphisms within the same target sequence.

2. Tiling Arrays

The polymorphisms can also be identified by hybridization to nucleic acid arrays, some example of which are described by WO 95/11995 (incorporated by reference in its entirety for all purposes). One form of such arrays is described in the Examples section in connection with de novo identification of polymorphisms. The same array or a different array can be used for analysis of characterized polymorphisms. WO 95/11995 also describes subarrays that are optimized for detection of a variant forms of a precharacterized polymorphism. Such a subarray contains probes designed to be complementary to a second reference sequence, which is an allelic variant of the first reference sequence. The second group of probes is designed by the same principles as described in the Examples except that the probes exhibit complementarily to the second reference sequence. The inclusion of a second group (or further groups) can be particular useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases).

3. Allele-Specific Primers

An allele-specific primer hybridizes to a site on target DNA overlapping a polymorphism and only primes amplification of an allelic form to which the primer exhibits perfect complementarily. See Gibbs, Nucleic Acid Res. 17, 2427-2448 (1989). This primer is used in conjunction with a second primer which hybridizes at a distal site. Amplification proceeds from the two primers leading to a

detectable product signifying the particular allelic form is present. A control is usually performed with a second pair of primers, one of which shows a single base mismatch at the polymorphic site and the other of which exhibits perfect

5   complementarily to a distal site. The single-base mismatch prevents amplification and no detectable product is formed. The method works best when the mismatch is included in the 3'-most position of the oligonucleotide aligned with the polymorphism because this position is most destabilizing to

10   elongation from the primer. *See, e.g.,* WO 93/22456.

          4.   Direct-Sequencing

          The direct analysis of the sequence of polymorphisms of the present invention can be accomplished using either the dideoxy chain termination method or the Maxam Gilbert method

15   (see Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd Ed., CSHP, New York 1989); Zyskind et al., *Recombinant DNA Laboratory Manual*, (Acad. Press, 1988)).

          5.   Denaturing Gradient Gel Electrophoresis

          Amplification products generated using the

20   polymerase chain reaction can be analyzed by the use of denaturing gradient gel electrophoresis. Different alleles can be identified based on the different sequence-dependent melting properties and electrophoretic migration of DNA in solution. Erlich, ed., *PCR Technology, Principles and*

25   *Applications for DNA Amplification,* (W.H. Freeman and Co, New York, 1992), Chapter 7.

          6.   Single-Strand Conformation Polymorphism Analysis

          Alleles of target sequences can be differentiated

30   using single-strand conformation polymorphism analysis, which identifies base differences by alteration in electrophoretic migration of single stranded PCR products, as described in Orita et al., *Proc. Nat. Acad. Sci.* 86, 2766-2770 (1989). Amplified PCR products can be generated as described above,

35   and heated or otherwise denatured, to form single stranded amplification products. Single-stranded nucleic acids may refold or form secondary structures which are partially dependent on the base sequence. The different electrophoretic

mobilities of single-stranded amplification products can be related to base-sequence difference between alleles of target sequences.

5 III. Methods of Use

After determining polymorphic form(s) present in an individual at one or more polymorphic sites, this information can be used in a number of methods.

10   A.   Forensics

Determination of which polymorphic forms occupy a set of polymorphic sites in an individual identifies a set of polymorphic forms that distinguishes the individual. *See generally* National Research Council, *The Evaluation of*

15 *Forensic DNA Evidence* (Eds. Pollard et al., National Academy Press, DC, 1996). The more sites that are analyzed the lower the probability that the set of polymorphic forms in one individual is the same as that in an unrelated individual. Preferably, if multiple sites are analyzed, the sites are

20 unlinked. Thus, polymorphisms of the invention are often used in conjunction with polymorphisms in distal genes. Preferred polymorphisms for use in forensics are diallelic because the population frequencies of two polymorphic forms can usually be determined with greater accuracy than those of multiple

25 polymorphic forms at multi-allelic loci.

The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue

30 sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was

35 not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been

determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance.

5

B.   Correlation of Polymorphisms with Phenotypic Traits

The polymorphisms of the invention may contribute to the phenotype of an organism in different ways. Some polymorphisms occur within a protein coding sequence and contribute to phenotype by affecting protein structure. The

10   effect may be neutral, beneficial or detrimental, or both beneficial and detrimental, depending on the circumstances. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on replication,

15   transcription, and translation. A single polymorphism may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by polymorphisms in different genes. Further, some polymorphisms predispose an individual to a distinct mutation that is causally related to a certain

20   phenotype.

Some disease traits are already known to be mitochondrially inherited. Some such disease traits result, at least in part, from stop codons in structural genes. Such mutations have been mapped and associated with diseases, such

25   as Leber's hereditary optic neuropathy, neurogenic muscular weakness, ataxia and retinitis pigmentosa. Other mutations (nucleotide substitutions) occur in tRNA coding sequences, and presumably cause conformational defects in transcribed tRNA molecules. Such mutations have also been mapped and

30   associated with diseases such as Myoclonic Epilepsy and Ragged Red Fiber Disease. *See* Wallace, *Ann. Rev. Biochem.* 61, 1175- 1212 (1992) (incorporated by reference in its entirety for all purposes).

Other genetic diseases having hitherto unmapped

35   genetic component(s) may also result in part from variations in mitochondrial DNA. Candidate diseases include, e.g., agammaglobulimenia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease,

familial hypercholesterolemia, polycystic kidney disease, hereditary spherocytosis, von Willebrand's disease, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, osteogenesis

5   imperfecta, and acute intermittent porphyria).

Other phenotypic traits that may derive, at least in part, from variations in mitochondrial DNA include symptoms of, or susceptibility to, multifactorial diseases of which a component is or may be genetic, such as Alzheimer's disease,

10   autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non-independent), systemic lupus erythematosus and Graves

15   disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, leukemia, liver, lung, oral cavity, ovary, pancreas, prostate, skin, stomach and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness,

20   obesity), strength, speed, endurance, fertility, and susceptibility or receptivity to particular drugs or therapeutic treatments.

Correlation is performed for a population of individuals who have been tested for the presence or absence

25   of a phenotypic trait of interest and for polymorphic markers sets. To perform such analysis, the presence or absence of a set of polymorphisms (i.e. a polymorphic set) is determined for a set of the individuals, some of whom exhibit a particular trait, and some of whom exhibit lack of the trait.

30   The alleles of each polymorphism of the set are then reviewed to determine whether the presence or absence of a particular allele is associated with the trait of interest. Correlation can be performed by standard statistical methods such as a κ-squared test and statistically significant correlations

35   between polymorphic form(s) and phenotypic characteristics are noted. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined

presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased milk production in a mother after birth of a child.

Such correlations can be exploited in several ways. In the case of a strong correlation between a set of one or more polymorphic forms and a disease for which treatment is available, detection of the polymorphic form set in a human or animal patient may justify immediate administration of treatment, or at least the institution of regular monitoring of the patient, before irretrievable damage has occurred. Treatments include metabolic replacement and gene therapy. *See* Wallace, *supra*. Detection of a polymorphic form correlated with serious disease in a couple contemplating a family may also be valuable to the couple in their reproductive decisions. For example, the female partner might elect to undergo *in vitro* fertilization to avoid the possibility of transmitting such a polymorphism from her husband to her offspring. In the case of a weaker, but still statistically significant correlation between a polymorphic set and human disease, immediate therapeutic intervention or monitoring may not be justified. Nevertheless, the patient can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little cost to the patient but confer potential benefits in reducing the risk of conditions to which the patient may have increased susceptibility by virtue of variant alleles. Identification of a polymorphic set in a patient correlated with enhanced receptiveness to one of several treatment regimes for a disease indicates that this treatment regime should be followed.

As an example of how mitochondrial polymorphisms can be correlated with phenotypic traits, Beitz et al., US 5,292,639 discuss use of bovine mitochondrial polymorphisms in a breeding program to improve milk production in cows. To evaluate the effect of mtDNA D-loop sequence polymorphism on milk production, each cow was assigned a value of 1 if variant or 0 if wildtype with respect to a prototypical mitochondrial DNA sequence at each of 17 locations considered. Each

production trait was analyzed individually with the following
animal model:

$$Y_{ijkpn} = \mu + YS_i + P_j + X_k + \beta_1 + \dots \beta_{17} + PE_n + a_n + e_p$$

where Y_{ijknp} is the milk, fat, fat percentage, SNF, SNF

5   percentage, energy concentration, or lactation energy record;
μ is an overall mean; YS_i is the effect common to all cows
calving in year-season; X_k is the effect common to cows in
either the high or average selection line; β_1 to β_{17} are the
binomial regressions of production record on mtDNA D-loop

10 sequence polymorphisms; PE_n is permanent environmental effect
common to all records of cow n; a_n is effect of animal n and
is composed of the additive genetic contribution of sire and
dam breeding values and a Mendelian sampling effect; and e_p is
a random residual. It was found that eleven of seventeen

15 polymorphisms tested influenced at least one production trait.
Bovines having the best polymorphic forms for milk production
at these eleven loci are used as parents for breeding the next
generation of the herd.

20 IV.   <u>Modified Polypeptides and Gene Sequences</u>

The invention further provides variant forms of
nucleic acids and corresponding proteins. The nucleic acids
comprise a segment of at least 10 bases of the nucleotide
sequence of Anderson, supra, except at one of the polymorphic

25 positions described in Table 1, column 1, in which the
polymorphic position is occupied by a different base than
described by Anderson, supra. Preferably, the different base
is one of the alternative bases for that position shown in
Table 1. Some nucleic acid encode full-length variant forms

30 of proteins. Similarly, variant proteins have the
prototypical amino acid sequences of proteins encoded by the
nucleic acid sequence of Anderson, supra, (read in-frame)
except at an amino acid encoded by a codon including one of
the polymorphic positions shown in Table 1. That position is

35 occupied by a different amino acid than described by Anderson,
supra, and preferably an amino acid encoded by a corresponding
codon of any of the alternative forms for the position shown
in the Table.

Variant genes can be expressed in an expression vector in which a variant gene is operably linked to a native or other promoter. Usually, the promoter is a eukaryotic promoter for expression in a mammalian cell. The transcription regulation

5   sequences typically include a heterologous promoter and optionally an enhancer which is recognized by the host. The selection of an appropriate promoter, for example trp, lac, phage promoters, glycolytic enzyme promoters and tRNA promoters, depends on the host selected. Commercially

10   available expression vectors can be used. Vectors can include host-recognized replication systems, amplifiable genes, selectable markers, host sequences useful for insertion into the host genome, and the like.

The means of introducing the expression construct

15   into a host cell varies depending upon the particular construction and the target host. Suitable means include fusion, conjugation, transfection, transduction, electroporation or injection, as described in Sambrook, *supra*. A wide variety of host cells can be employed for expression of

20   the variant gene, both prokaryotic and eukaryotic. Suitable host cells include bacteria such as *E. coli*, yeast, filamentous fungi, insect cells, mammalian cells, typically immortalized, *e.g.*, mouse, CHO, human and monkey cell lines and derivatives thereof. Preferred host cells are able to

25   process the variant gene product to produce an appropriate mature polypeptide. Processing includes glycosylation, ubiquitination, disulfide bond formation, general post-translational modification, and the like.

The protein may be isolated by conventional means of

30   protein biochemistry and purification to obtain a substantially pure product, *i.e.*, 80, 95 or 99% free of cell component contaminants, as described in Jacoby, *Methods in Enzymology* Volume 104, Academic Press, New York (1984); Scopes, *Protein Purification, Principles and Practice*, 2nd

35   Edition, Springer-Verlag, New York (1987); and Deutscher (ed), *Guide to Protein Purification, Methods in Enzymology,* Vol. 182 (1990). If the protein is secreted, it can be isolated from the supernatant in which the host cell is grown. If not

secreted, the protein can be isolated from a lysate of the host cells.

The invention further provides transgenic nonhuman animals capable of expressing an exogenous variant gene and/or having one or both alleles of an endogenous variant gene inactivated. Expression of an exogenous variant gene is usually achieved by operably linking the gene to a promoter and optionally an enhancer, and microinjecting the construct into a zygote. See Hogan et al., "Manipulating the Mouse Embryo, A Laboratory Manual," Cold Spring Harbor Laboratory. Inactivation of endogenous variant genes can be achieved by forming a transgene in which a cloned variant gene is inactivated by insertion of a positive selection marker. See Capecchi, *Science* 244, 1288-1292 (1989). The transgene is then introduced into an embryonic stem cell, where it undergoes homologous recombination with an endogenous variant gene. Mice and other rodents are preferred animals. Such animals provide useful drug screening systems.

In addition to substantially full-length polypeptides expressed by variant genes, the present invention includes biologically active fragments of the polypeptides, or analogs thereof, including organic molecules which simulate the interactions of the peptides. Biologically active fragments include any portion of the full-length polypeptide which confers a biological function on the variant gene product, including ligand binding, and antibody binding. Ligand binding includes binding by nucleic acids, proteins or polypeptides, small biologically active molecules, or large cellular structures.

Polyclonal and/or monoclonal antibodies that specifically bind to variant gene products but not to corresponding prototypical gene products are also provided. Antibodies can be made by injecting mice or other animals with the variant gene product or synthetic peptide fragments thereof. Monoclonal antibodies are screened as are described, for example, in Harlow & Lane, *Antibodies, A Laboratory Manual*, Cold Spring Harbor Press, New York (1988); Goding, *Monoclonal antibodies, Principles and Practice* (2d ed.)

Academic Press, New York (1986). Monoclonal antibodies are tested for specific immunoreactivity with a variant gene product and lack of immunoreactivity to the corresponding prototypical gene product. These antibodies are useful in
5   diagnostic assays for detection of the variant form, or as an active ingredient in a pharmaceutical composition.

V.   Kits

      The invention further provides kits comprising at
10 least one allele-specific oligonucleotide as described above. Often, the kits contain one or more pairs of allele-specific oligonucleotides hybridizing to different forms of a polymorphism. In some kits, the allele-specific oligonucleotides are provided immobilized to a substrate. For
15 example, the same substrate can comprise allele-specific oligonucleotide probes for detecting at least 10, 100 or all of the polymorphisms shown in Table 1. Optional additional components of the kit include, for example, restriction enzymes, reverse-transcriptase or polymerase, the substrate
20 nucleoside triphosphates, means used to label (for example, an avidin-enzyme conjugate and enzyme substrate and chromogen if the label is biotin), and the appropriate buffers for reverse transcription, PCR, or hybridization reactions. Usually, the kit also contains instructions for carrying out the methods.
25

EXAMPLES

      An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect
30 differences between the target and a reference sequence. Many different arrays can be designed for these purposes. One such design, termed a 4L tiled array, is depicted schematically in Fig. 1A. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. By this
35 approach, a nucleic acid target of length L can be scanned for mutations with a tiled array containing 4L probes. For example, to query the 16,569 bp of human mitochondrial DNA

(mtDNA), only 66,276 probes of the possible $\sim 10^9$ 15-mers need to be used.

The use of a tiled array of probes to read a target sequence is illustrated in Fig. 1C. A tiled array of 15-mers varied at position 7 from the 3' end ($P^{15,7}$) was designed and synthesized for mt1, a cloned sequence containing 1,311 bp spanning the control region of mtDNA (Anderson, et al., *Nature* **290**, 457-465 (1981); Greenberg et al., *Gene* **21**, 33-49 (1983); Aquardo & Greenberg, *Genetics* **103**, 287-312 (1983); Cann et al. *Genetics* **106**, 479-499 (1984).

The mt1 and mt2 sequences were cloned from amplified genomic DNA extracted from hair roots (Gill et al. *Nature* **318**, 577-579 (1985); Saiki, et al., *Science* **239**, 487-491 (1988)). The clones were sequenced conventionally. Cloning was performed only to provide a set of pure reference samples of known sequence. To provide templates for fluorescent labeling, DNA was reamplified from the clones using primers bearing bacteriophage T3 and T7 RNA polymerase promoter sequences (bold; mtDNA sequences uppercase): L15935-T3, 5'ctcgg**aattaaccctcactaaagg**AAACCTTTTTCCAAGGA and H667-T7, 5'**taatacgactcactataggg**agAGGCTAGGACCAAACCTATT.

The upper panel of Fig. 1C shows a portion of the fluorescence image of an array hybridized with fluorescein-labeled mt1 RNA. Labeled RNAs from the two complementary 2mtDNA strands (designated L and H) were transcribed in separate reactions from a promoter-tagged PCR product. Each 10 μl reaction contained 1.5 mM each of ATP, CTP, GTP and UTP, 0.24 mM fluorescein-12-CTP (Du Pont), 0.24 mM fluorescein-12-UTP (Boehringer Mannheim), $\sim$ 1 to 5 nM (1.5 μl) crude unpurified 1.3 kb PCR product and 1 unit/ml T3 or T7 RNA polymerase (Promega), in a reaction buffer supplied with the enzyme. The reaction was carried out at 37°C for 1 to 2 h. RNA was fragmented to an average size of < 100 nt by adjusting the solution to 30 mM $MgCl_2$, by the addition of 1 M $MgCl_2$, and heating at 94°C for 40 minutes. Fragmentation improved the uniformity and specificity of hybridization. The extent of fragmentation is dependent on magnesium ion concentration [Huff et al., *Biochem.* **3**, 501-506 (1964); Butzow

et al. *Biopolymers* **3**, 95-107 (1965)]. Good hybridization
results have been obtained with both DNA and RNA targets
prepared using a variety of labelling schemes, including
incorporation of fluorescent and biotinylated dNTPs by DNA
5   polymerases, incorporation of dye-labelled primers during PCR,
ligation of labelled oligonucleotides to fragmented RNA, and
direct labelling by photo-crosslinking a psoralen derivative
of biotin directly to fragmented nucleic acids.

The base sequence can be read by comparing the
10   intensities of the four probes within each column. For
example, the column for position 16,493 consists of the four
probes, 3'_TGACAT**A**GGCTGTAG, 3'_TGACAT**C**GGCTGTAG,
3'_TGACAT**G**GGCTGTAG, and 3'_TGACAT**T**GGCTGTAG. The probe with
the strongest signal is the probe with the **A** substitution (**A**
15   49, **C** 8, **G** 15, **T** 8 counts, 2 counts background), identifying
the base at position 16,493 as U in the RNA transcript.
Continuing the process, the sequence at each position can be
read directly from the hybridization intensities.

The effect on the array hybridization pattern caused by a
20   single base change in the target is shown in Fig. 1B, and the
detection of a single-base polymorphism is shown in the lower
panel of Fig. 1C. The target here is mt2, which differs from
mt1 in this region by a T to C transition at position 16,493.
Accordingly, the probe with the **G** substitution (third row)
25   displays the strongest signal. Since the tiled array was
designed to mt1, the hybridization intensities of neighboring
probes that overlap 16,493 are also affected by the change in
target sequence. The hybridization signals of 60 probes (4 x
15) of the 15-mer tiled array are perturbed by a single-base
30   change in the target sequence. In the $P^{15,7}$ array, each probe
querying the 8 positions to the left and 6 positions to the
right of the polymorphism contain at least one mismatch to the
target. The result is a characteristic loss of signal or a
"footprint" for the probes flanking a mutation position. Of
35   the four probes querying each position, the loss of signal is
greatest for the one designed to match mt1. We denote the
subset of probes with zero mismatches to the reference
sequence as P^0.

A comparison of P^0 hybridization signals from a target to those from a reference is ideally obtained by hybridizing both samples to the same array. We therefore developed a two-color labeling and detection scheme in which the reference is labeled with phycoerythrin (red), and the target with fluorescein (green). The reference and unknown samples were labeled with biotin and fluorescein, respectively, in separate transcription reactions. Reactions were carried out as described above except that each contained 1.25 mM of ATP, CTP, GTP and UTP and 0.5 mM fluorescein-12-UTP or 0.25 mM biotin-16-UTP (Boehringer Mannheim). The two reactions were mixed in the ratio 1:5 (v/v) biotin:fluorescein and fragmented. Targets were diluted to a final concentration of ~ 100 to 1000 pM in 3 M TMACl [Melchior et al., *Proc. Natl. Acad. Sci. USA* 70, 298-302 (1973)], 10 mM Tris.Cl pH 8.0, 1mM EDTA, 0.005% Triton X-100, and 0.2 nM of a control oligonucleotide, 5' fluorescein- CTGAACGGTAGCATCTTGAC. Samples were denatured at 95°C for 5 min, chilled on ice for 5 min and equilibrated to 37°C. A volume of 180 μl of hybridization solution was then added to the flow cell [Lipshutz et al., Biotechniques **19**, 442-447 (1995)], and the chip incubated at 37°C for 3 h with rotation at 60 rpm. The chip was washed 6 times at RT with 6xSSPE, 0.005% Triton X-100. Phycoerythrin-conjugated streptavidin (2 μg/ml in 6xSSPE, 0.005 % Triton X-100) was added and incubation continued at RT for 5 min. The chip was washed again, and rscanned at a resolution of ~ 74 pixels per probe cell. Two scans were collected, a fluorescein scan using a 515 to 545 nm bandpass filter, and a phycoerythrin scan using a 560 nm longpass filter. Signals were separated to remove spectral overlap and average counts per cell determined.

By processing the reference and target together, experimental variability during the fragmentation, hybridization, washing, and detection steps is minimized or eliminated. In addition, during co-hybridization of the reference and target, competition for binding sites results in a slight improvement in mismatch discrimination. However, array hybridization is highly reproducible, and comparative analysis of data obtained

from separate but identically synthesized arrays is also effective.

The two-color approach was tested by analyzing a 2.5 kb region of mtDNA, that spans the tRNA^{Glu}, cytochrome b, tRNA^{Thr}, tRNA^{Pro}, control region and tRNA^{Phe} DNA sequences. Each 2.5 kb target sequence was PCR amplified directly from genomic DNA using the primer pair L14675-T3 (5'aattaaccctcactaaagggATTCTCGCACGGACTACAAC) and H667-T7. A p^{20,9} array (i.e., containing 20-mer probes varied at position 9 from the 3' end) was designed to match the mt1 target (i.e., p^0 sequence = mt1). The mt1 reference (red) and a polymorphic target sample (green) were pooled and hybridized simultaneously to the array. Differences between the target and reference sequences were identified by comparing the scaled red and green p^0 hybridization intensities. To scale the sample to the reference intensities, a histogram of the base 10 logarithm of the intensity ratios for each pair of probes was constructed. The histogram had a mesh size of 0.01, and was smoothed by replacing the value at each point with the average number of counts over a five-point window centered at that point. The highest value in the histogram was located, and the resulting intensity ratio was taken to be the most probable calibration coefficient.

The marked decrease in target hybridization intensity, over a span of ⁻20 nucleotides, is shown for a single base polymorphism at position 16,223 (Fig. 2A). The footprint is enlarged when two polymorphisms occur in close proximity (within ⁻20 nucleotides) as shown in Fig. 2B. When polymorphisms are clustered, the size of the footprint depends on the number of polymorphisms and their separation (Fig. 2C).

Polymorphisms, particularly multiple polymorphisms within a probe length, can be read most accurately by combining the results of basecalling and footprinting. Closely spaced multiple polymorphisms are easily identified by the presence of a large footprint (Fig. 2B and C) or by two or more basecalls differing from P^0 within a single probe span. Discrepancies between basecalling and footprint patterns are flagged for further analysis (for example, a P^0 footprint in

which no polymorphism is identified: such a pattern is typical
of a deletion). Thus, basecalls are valid only for unflagged
regions. In flagged regions, the presence of sequence
differences is detected, but no attempt is made to call the
5   sequence without further analysis.

Sequence analysis was carried out on the 2.5 kb target
from 12 samples. A total of 30,582 bp containing 180
substitutions relative to mt1 was analyzed. 98% of the
sequence was unambiguously assigned by a Bayesian basecalling
10   algorithm. Base identification was accomplished using a
Bayesian classification algorithm based on variable kernel
density estimation. The likelihood of each basecall
associated with a set of hybridization intensity values was
computed by comparing an unknown set of probes to a set of
15   example cases for which the correct basecall was known. The
resulting four likelihoods were then normalized so that they
summed to 1. Data from both strands were combined by averaging
the values. If the most likely basecall had an average
normalized likelihood greater than 0.6, it was called,
20   otherwise the base was called as an ambiguity. The example set
was derived from 2 different samples, ib013 and ief005, which
have a total of 35 substitutions relative to mt1, of which 19
are shared with the 12 samples analyzed and 16 are not.
Basecalling performance was not sensitive to the choice of
25   examples. Of this 98%, which contained both wild-type
sequence and a high proportion of single-base footprints such
as the example shown in Fig. 2A, 29,878 out of 29,879 bp were
called correctly.

To provide an independently determined reference
30   sequence, each 2.5 kb PCR amplicon was sequenced on both
strands by primer-directed fluorescent chain-terminator cycle
sequencing using an ABI 373A DNA sequencer, and assembled and
manually edited using Sequencer 3.0. The analysis presented
here assumes that the sequence amplified from genomic DNA is
35   essentially clonal [Monnat & Loeb, *Proc. Natl. Acad. Sci. USA*
82, 2895-2899 (1985)], and that its determination by gel-based
methods is correct. A common length polymorphism at positions
303-309 was not detected by hybridization under the conditions

used. It was excluded from analysis, and is not part of the
set of 180 polymorphisms discussed in the text. However,
polymorphisms at this site have previously been differentiated
by oligonucleotide hybridization [Stoneking et al., *Am. J.*
5   *Hum. Genet.* **48**, 370-382 (1991)].

The remaining 2%, that contained the multiple
substitution footprints (such as those shown in Fig. 2B and
2C), was flagged for further analysis. Of the 649 bp
comprising this 2%, 643 bp were located in or immediately
10   adjacent to footprints. The P^0 intensity footprints were
detected in the following way: the reference and sample
intensities were normalized (15), and R, the average of $\log_{10}$
(P^0 reference / P^0 sample) over a window of 5 positions,
centered at the base of interest, calculated for each position
15   in the sequence. Footprints were detected as regions having
at least 5 contiguous positions with a reference or sample
intensity at least 50 counts above background, and an R value
in the top 10th percentile for the experiment. At 205
polymorphic sites, where the sample was mismatched to P^0, the
20   mean R value was 1.01, with a standard deviation of 0.57. At
35,333 non polymorphic sites (i.e. where both reference and
sample had a perfect match to P^0) the mean value was -0.05,
with a standard deviation of 0.25.

In all, 179 out of the 180 polymorphisms were
25   unambiguously detected; 126 out of 127 were called correctly
in the unflagged regions; and 53 polymorphisms occurring in
the flagged regions were detected as footprints. There were
no false positive basecalls, and only one false positive
footprint. These figures can be considered to be "worst case"
30   estimates for the type of array and target used. The P^0
sequence represents a Caucasian haplotype, and our sample set
included 8 African samples having a large number of clustered
differences to P^0. Furthermore, the variation in the
hypervariable part of the control region is much higher than
35   for the rest of the mitochondrial genome and for nuclear genes
in general (Fig. 2A, B and C show comparisons to African
samples in this region).

We therefore designed a $P^{25,13}$ tiling array for the mitochondrial genome. The array contains a total of 136,528 synthesis cells, each ˜ 35 x 35 μm in size (Fig. 3). In addition to a 4L tiling across the genome, the array contains a set of probes representing a single base deletion at every position across the genome, and sets of probes designed to match a range of specific mtDNA haplotypes. Using long range PCR, the 16.6 kb mtDNA was amplified directly from genomic DNA samples. Long range PCR amplification was carried out on genomic DNA using Perkin Elmer GeneAmp XL PCR reagents according to the manufacturer's protocol. Primers were L14836-T3 (5'aattaaccctcactaaagggATGAAACTTCGGCTCACTCCTTGGC) and RH1066-T7 (5'taatacgactcactatagggaTTTCATCATGCGGAGATGTTGGATGG), based on RH 1066 [Cheng et al., *Nature Genetics* 7, 350-351 (1994)]. Each 100 μl reaction contained 0.2 μM of each primer and ˜ 10-50 ng total genomic DNA. Transcription reactions were carried out using Ambion MAXIscript kit using the manufacturer's protocol. The concentration of the 16.6 kb PCR template was ˜ 2 nM and the reaction was spiked with 0.2 mM biotin-16-UTP. Incubation was at 37°C for 2 h. Fragmentation and hybridization were as described above, except that 3.5 M TMACl and 5' biotin-CTGAACGGTAGCATCTTGAC were used in the hybridization buffer, which also contained 100 μg/ml fragmented baker's yeast RNA (Sigma); and hybridization was carried out at 40°C for 4 h. Labeled RNA targets were prepared by in vitro transcription and hybridized to the array. Genomic hybridization patterns were imaged in less than 10 minutes by a high resolution confocal microscope.

A custom telocentric objective lens with a numerical aperture of 0.25 focuses 5 mW of 488 nm argon laser light to a 3 μm diameter spot, which is scanned by a galvanometer mirror across a 14 mm field at 30 lines/sec. Fluorescence collected by the objective is descanned by the galvanometer mirror, filtered by a dichroic beamsplitter (555 nm) and a bandpass filter (555-607 nm), focused onto a confocal pinhole, and detected by a photomultiplier. Photomultiplier output is digitized to 12 bits. A 4096 x 4096 pixel image is obtained in

less than 3 min. Pixel size is 3.4 μm. The data from four
sequential scans were summed to improve signal to noise.

Figure 3B shows the hybridization pattern of a 16.6 kb
target to the mitochondrial genome chip. Although there are
some regions of low intensity, most of the 25-mer array
hybridized efficiently: simply by identifying the highest
intensity in each column of 4 substitution probes, 99.0 % of
the mt3 sequence could be read correctly (P^0 sequence = mt3).
The array was used to successfully detect three
disease-causing mutations in a mitochondrial DNA sample from a
patient suffering from Leber's hereditary optic neuropathy
(Brown et al., *FASEB J.* **6**, 2791-2799 (1992)) (Fig. 3C).
Furthermore, 7 errors and undetected polymorphisms from the
gel-based sequence were identified.

We then hybridized 10 African genomes to the array and
unambiguously identified 505 polymorphisms. These were
polymorphisms that could be clearly read and for which a
confirmatory footprint was detected automatically. For the 10
samples, the 2.5 kb cytochrome b and control region sequences
were known. No false positives were detected in the ¯ 25 kb
of sequence checked in this way. Additional clustered
polymorphisms were detected by the presence of footprints, but
not read directly.

The throughput of the system we describe is 5 chips per
hour. Thus, 50 genomes can be read by hybridization in the
time it takes to read 2 genomes conventionally. Furthermore,
there are drastic reductions in sample preparation needs,
because the entire genome is labeled in a single reaction, at
a cost similar to that for a single sequencing reaction.
Also, sequence reading at the level of data analysis is
automated: the sequences can be read in a matter of minutes.
No analytical separations or gel preparation are needed, which
contributes to the speed of the experiment. Finally, a clear
advantage to the approach we describe is that it is highly
scalable. The cost, effort and time required to analyze the
entire 16.6 kb mtDNA in a single experiment is virtually
identical to that required to read 2.5 kb.

From the foregoing, it is apparent that the
invention includes a number of general uses that can be
expressed concisely as follows. The invention provides for
the use of any of the nucleic acid segments described above in

5   the diagnosis or monitoring of diseases, such as Alzheimer's
disease, cancer, inflammation, heart disease, diseases of the
CNS, and susceptibility to infection by microorganisms. The
invention further provides for the use of any of the nucleic
acid segments in the manufacture of a medicament for the

10   treatment or prophylaxis of such diseases. The invention
further provides for the use of any of the DNA segments as a
pharmaceutical.

Although the foregoing invention has been described in
detail for purposes of clarity of understanding, it will be

15   obvious that certain modifications may be practiced within the
scope of the appended claims. All publications and patent
documents cited above are hereby incorporated by reference in
their entirety for all purposes to the same extent as if each
were so individually denoted.

20

Table 1

pos	base	asn_base	HA001_v_mt2	HA002_v_mt2	HA004_v_mt2	HA007_v_mt2	IB013_v_mt2	IEF002_v_mt2	IEF005_v_mt2	IEF007_v_mt2	IEF011_v_mt2	YR019
63	C	c	C	C	T	C	C	C	C	C	C	C
72	G	a	G	G	A	G	G	G	G	G	G	G
92	A	a	A	g	G	G	A	A	A	A	A	A
181	C	c	T	T	c	T	c	t	C	C	C	C
194	T	t	c	C	t	C	t	c	T	C	C	C
234	G	a	A	A	g	A	a	A	A	A	A	A
246	G	g	G	G	A	G	a	G	G	G	G	G
296	A	a	A	A	A	A	G	A	A	A	A	A
315	G	g	G	G	G	G	A	G	G	G	G	G
324	C	c	C	T	C	T	C	C	C	C	C	C
662	G	a	A	A	A	A	A	A	A	A	A	A
679	T	t	T	C	T	C	T	T	T	T	T	T
708	G	g	A	A	G	A	G	A	G	G	G	G
768	G	g	A	a	a	a	a	A	G	A	A	a
824	T	t	T	T	A	T	A	T	T	T	T	T
1017	G	g	A	A	A	A	A	A	G	A	A	A
1039	T	t	T	T	T	C	T	T	T	T	T	T
1047	C	c	C	C	T	C	C	C	C	C	C	C
1441	G	g	A	A	G	A	G	A	G	G	G	G
1705	C	c	T	C	C	C	C	T	C	C	C	C
1735	G	a	A	A	A	A	A	A	A	A	A	A
2357	A	a	G	A	A	A	A	G	A	A	A	A
2415	T	t	t	t	T	t	T	A	T	n	t	t
2757	G	g	G	G	A	G	A	G	G	G	G	G

Table 1 continued

2788	C	c	C	C	C	C	C	C	C	T	T	T
2884	T	t	T	T	C	T	C	T	T	T	T	T
3009	G	g	G	G	G	G	G	G	G	G	G	A
3199	T	t	T	A	T	A	T	T	T	T	T	T
3335	T	t	T	T	T	T	T	T	T	T	C	T
3336	G	g	G	A	G	G	G	G	G	G	G	G
3449	C	c	C	C	C	C	C	C	T	C	C	C
3515	C	c	C	C	A	C	C	C	C	C	C	C
3593	C	c	T	T	T	t	c	T	C	T	T	T
3647	C	c	T	C	C	C	C	C	C	C	C	C
3665	G	g	G	G	G	G	A	G	G	G	G	G
3744	G	g	G	A	G	G	G	G	G	G	G	G
3865	T	t	T	T	C	T	T	T	T	T	T	T
3917	G	g	G	G	G	G	G	G	G	A	A	G
4103	A	a	G	G	G	G	G	G	A	G	G	G
4157	A	a	G	A	A	A	A	G	A	A	A	A
4247	C	t	T	T	T	T	T	T	T	T	T	T
4311	C	c	C	C	T	C	C	C	C	C	C	C
4585	T	t	T	T	C	T	T	T	T	T	T	T
4766	A	a	G	A	A	A	A	G	A	A	A	A
4823	G	a	A	A	A	A	A	A	A	A	A	A
4966	T	t	T	T	T	C	T	T	T	T	T	T
5026	C	c	T	C	C	C	C	T	C	C	C	C
5095	T	t	T	T	C	T	T	T	T	T	T	T
5146	G	g	G	G	G	G	G	G	G	G	A	G
5230	G	g	G	G	A	G	G	G	G	G	G	G
5251	G	g	G	G	G	G	G	G	G	G	G	A
5284	A	a	A	A	A	A	A	A	A	g	G	A
5330	C	c	A	C	C	C	C	A	C	C	C	C
5459	G	g	G	G	A	G	G	G	G	G	G	G
5602	C	c	C	C	T	C	C	C	C	C	C	C
5772	G	g	G	G	G	G	G	G	A	G	G	G
5813	T	t	C	T	T	T	T	C	T	T	T	T
5822	A	a	A	A	A	G	A	A	A	A	A	A
5911	C	c	T	C	C	C	C	C	C	C	C	C
6070	T	t	T	T	T	T	C	T	T	T	T	T
6149	G	g	G	G	G	G	A	G	G	G	G	G
6236	C	c	C	C	C	C	C	C	T	C	C	C
6252	T	t	T	T	T	T	C	T	T	T	T	T
6526	G	a	A	A	A	A	A	A	A	A	A	A
6613	T	t	C	T	T	T	T	C	T	T	T	T
6662	A	a	A	A	A	A	A	A	A	A	A	G
6712	C	c	T	C	C	C	C	T	C	C	C	C
6805	A	a	G	A	A	A	A	G	A	A	A	A
6826	T	t	T	C	T	T	T	T	T	T	T	T
6874	C	c	C	C	C	C	C	C	C	A	C	C
7145	A	a	A	A	G	A	G	A	A	A	A	A
7174	T	t	T	T	T	T	T	T	T	C	C	C
7255	C	c	T	T	T	T	T	T	C	T	T	T
7273	C	c	C	C	C	C	C	C	C	T	T	T
7388	T	t	T	T	T	T	C	T	T	T	T	T
7517	A	a	A	n	A	G	a	a	A	N	A	A
7520	G	g	A	c	A	G	g	g	G	A	A	A

Table 1 continued

7623	T	t	A	A	T	A	T	A	T	T	T	T
7770	A	a	A	A	A	A	A	A	A	G	G	g
8026	A	g	G	G	G	G	A	G	G	G	G	G
8079	C	c	T	C	C	C	C	T	C	C	C	C
8205	G	g	A	A	G	A	G	A	G	A	A	A
8386	G	g	A	G	G	G	G	A	G	G	G	G
8427	C	c	C	C	T	C	C	C	C	C	C	C
8467	C	c	C	C	T	C	T	C	C	C	c	C
8502	T	t	C	T	T	c	T	C	T	T	T	T
8565	A	a	A	A	G	A	A	A	A	A	A	A
8654	C	c	C	C	T	C	T	C	C	C	C	C
8700	A	a	G	G	G	G	G	G	G	G	G	G
8783	A	a	A	A	A	A	G	A	A	A	A	A
8793	T	c	C	C	C	C	C	C	C	C	C	C
8853	G	g	G	G	G	G	G	G	G	G	G	A
8876	T	t	T	T	T	T	C	T	T	T	T	T
9041	C	c	C	C	T	C	C	C	C	C	C	C
9071	A	a	A	A	A	A	G	A	A	A	A	A
9220	A	a	G	G	A	G	A	G	A	G	G	G
9346	A	a	A	A	G	A	A	A	A	A	A	A
9448	C	c	C	C	C	C	C	C	T	C	C	C
9817	C	c	C	C	T	C	C	C	C	C	C	C
10030	T	t	T	T	T	T	C	T	T	T	T	T
10114	T	t	C	C	T	c	T	C	T	C	C	T
10320	T	t	T	T	T	T	C	T	T	T	T	T
10372	G	g	G	G	G	G	G	G	A	G	G	G
10585	G	g	G	G	g	G	A	G	G	G	G	G
10663	C	c	C	C	T	C	C	C	C	C	C	C
10687	G	g	G	G	A	G	A	G	G	G	G	G
10792	C	c	C	C	C	C	T	C	C	C	C	C
10809	T	t	T	T	C	T	C	T	T	T	T	T
10827	T	t	T	T	T	T	C	T	T	T	T	T
10872	T	t	C	C	C	C	C	C	C	C	C	C
10914	T	t	T	T	C	T	T	T	T	T	t	T
11163	A	a	A	A	A	A	G	A	A	A	A	A
11175	A	g	G	G	A	G	G	G	G	G	G	G
11640	A	a	A	A	G	A	A	A	A	A	A	A
11653	A	a	A	A	A	A	G	A	A	A	A	A
11799	A	a	A	A	A	A	A	A	G	A	A	A
11913	G	g	G	G	A	G	G	G	G	A	A	A
11943	T	t	C	C	T	C	T	C	T	C	C	C
12006	A	g	G	G	A	G	G	G	G	G	G	G
12048	C	c	C	C	C	C	T	C	C	C	C	C
12114	C	c	C	C	C	C	T	C	C	C	C	C
12235	G	g	A	A	G	A	G	A	G	G	G	G
12353	T	t	T	T	T	T	T	T	T	T	T	C
12476	T	t	T	T	T	T	C	T	T	T	T	T
12719	A	a	A	A	G	A	A	A	A	A	A	A
12776	A	a	G	A	A	A	A	A	A	A	A	A
12809	A	a	A	A	A	A	G	A	A	A	A	A
12947	A	a	G	A	A	A	A	G	A	A	A	A
13104	A	a	A	A	G	A	G	A	G	A	A	A
13148	A	a	A	A	A	A	G	A	A	A	A	A

Table 1 continued

13183	T	t	T	T	T	C	T	T	T	T	T	T
13202	A	a	A	A	A	G	A	A	A	A	A	A
13275	A	a	A	A	G	A	A	A	A	A	A	A
13484	A	a	A	A	A	A	G	A	A	A	A	A
13505	C	c	C	C	T	C	t	C	C	C	C	C
13589	G	g	a	A	G	A	G	A	G	A	A	A
13649	C	c	t	t	t	t	T	T	C	T	T	t
13788	T	t	T	T	T	T	C	T	T	T	t	T
13802	A	a	A	A	A	A	A	A	A	g	G	g
13913	C	c	C	C	C	C	C	C	A	C	C	C
13957	G	g	G	C	G	C	G	G	G	G	G	G
13965	A	a	G	A	A	A	A	G	A	A	A	A
13999	T	t	T	T	T	T	A	T	T	T	T	T
14058	A	a	G	A	A	A	A	G	A	A	A	A
14177	T	t	T	T	T	T	C	T	T	T	T	T
14307	T	t	T	T	C	T	T	T	T	T	T	T
14406	C	c	T	C	C	C	C	T	C	C	C	C
14559	G	g	G	G	G	G	A	G	G	G	G	G
14565	A	a	A	A	A	A	A	A	A	G	G	G
14910	C	c	C	C	C	C	T	C	C	C	C	C
15042	G	g	G	G	G	A	G	G	G	G	G	G
15066	T	t	T	T	T	T	C	T	T	T	T	T
15109	G	g	A	A	G	A	G	A	G	G	G	G
15118	G	g	G	G	G	G	G	G	G	G	G	A
15135	C	c	C	C	T	C	C	C	C	C	C	C
15216	G	g	A	A	G	A	G	A	G	G	G	G
15243	A	a	A	A	A	A	A	A	A	G	G	A
15300	G	g	A	A	G	A	G	A	a	A	A	A
15430	G	g	G	G	A	G	G	G	G	G	G	G
15628	T	t	T	T	T	T	T	T	T	C	C	T
15733	G	g	G	G	G	G	G	G	G	G	G	A
15783	T	t	T	T	T	T	T	T	T	C	C	C
15848	C	c	C	T	C	T	C	C	C	C	C	C
15882	G	g	G	G	G	G	G	G	A	G	G	G
15901	A	a	A	A	A	A	A	A	G	A	A	A
16092	T	t	T	T	T	T	C	T	T	T	T	T
16110	T	c	C	C	C	C	C	C	C	C	C	C
16113	C	c	A	C	C	C	C	A	C	C	C	C
16123	T	t	t	T	T	T	T	T	C	T	T	T
16125	T	t	t	T	T	T	T	T	T	T	C	T
16128	G	g	A	G	A	G	G	G	G	G	G	G
16144	G	g	G	G	G	G	G	A	G	G	G	G
16147	C	c	C	C	T	C	C	C	C	C	C	C
16263	C	c	C	C	C	T	c	C	C	C	c	C
16277	C	c	T	T	C	T	t	T	T	t	T	T
16289	T	c	C	C	C	C	t	C	C	t	c	c
16317	A	a	A	G	a	A	A	A	A	A	A	A
16318	A	g	G	G	g	G	G	G	G	G	G	G
16353	C	c	T	C	C	C	C	T	C	C	C	C
16361	C	t	T	T	C	T	T	T	C	T	T	T
16389	G	g	A	a	G	a	G	A	G	A	A	A
16518	T	t	T	T	C	T	c	T	C	C	C	C

WHAT IS CLAIMED IS:

1   1.   A segment of human mitochondrial DNA or RNA of
2 between 10 and 100 bases including any one of the polymorphic
3 sites shown in Table 1, or the complement of the segment.

1   2.   The segment of claim 1, wherein the polymorphic
2 form occupying the polymorphic site is listed in Table 1,
3 column 3.

1   3.   The segment of claim 1, wherein the polymorphic
2 form occupying the polymorphic site is an alternative form
3 listed in Table 1, column 2, or 4-11.

1   4.   An allele-specific oligonucleotide that
2 hybridizes to a segment of human mitochondrial nucleic acid or
3 its complement including a polymorphic site shown in Table 1,
4 column 1.

1   5.   The allele-specific oligonucleotide of claim 10
2 that is probe.

1   6.   The allele-specific oligonucleotide of claim
2 10, wherein a central position of the probe aligns with the
3 polymorphic site of the fragment.

1   7.   The allele-specific oligonucleotide of claim 10
2 that is a primer.

1   8.   The allele-specific oligonucleotide of claim
2 13, wherein the 3' end of the primer aligns with the
3 polymorphic site of the fragment.

1   9.   An isolated nucleic acid comprising a segment
2 of the human mitochondrial sequence described by Anderson et
3 al., *Nature* 290, 457-465 (1981), or the complement thereof,
4 including a polymorphic site shown in Table 1, column 1,
5 wherein the polymorphic site within the segment is occupied by

6 a base other than the base shown in Table 1, column 3 ("asn
7 base").

1   10. A method of analyzing a nucleic acid,
2 comprising:
3   obtaining the nucleic acid from an individual; and
4   determining a base occupying any one of the polymorphic
5 sites shown in Table 1.

ABSTRACT

The invention provides novel human mitochondrial polymorphisms, and probes and primers for detecting the same. Detection of such polymorphisms is useful in a variety of
5   fields such as forensic analysis, epidemiology and preventive medicine.

J:\16528\250\app-1

DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I declare that:

My residence, post office address and citizenship are as stated below next to my name; I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural inventors are named below) of the subject matter which is claimed and for which a patent is sought on the invention entitled: **POLYMORPHISMS IN HUMAN MITOCHONDRIAL DNA** the specification of which __X__ is attached hereto or ___ was filed on ___ as Application No. ___ and was amended on ___ (if applicable).

I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above. I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, Section 1.56. I claim foreign priority benefits under Title 35, United States Code, Section 119 of any foreign applications(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)

Country	Application No.	Date of Filing	Priority Claimed Under 35 USC 119
			Yes _ No _
			Yes _ No _

I hereby claim the benefit under Title 35, United States Code §119(e) of any United States provisional application(s) listed below:

Application No.	Filing Date
60/017,203	May 16, 1996
60/024,206	August 20, 1996

I claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, Section 1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

Application No.	Date of Filing	Status
		_ Patented _ Pending _ Abandoned
		_ Patented _ Pending _ Abandoned

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith.

joe Liebeschuetz, Reg. No. 37,505
William M. Smith, Reg. No. 30,223
Renee A. Fitts, Reg. No. 35,136

Send Correspondence to: **William M. Smith** **TOWNSEND and TOWNSEND and CREW LLP** **Two Embarcadero Center, 8th Floor** **San Francisco, CA 94111-3834**	Direct Telephone Calls to: (Name, Reg. No., Telephone No.) Name: Joe Liebeschuetz Reg. No.: 37,505 Telephone: 415-326-2400

Full Name of Inventor 1	Last Name **CHEE**	First Name **Mark**	Middle Name or Initial	
Residence & Citizenship	City **Palo Alto**	State/Foreign Country **California**	Country of Citizenship **Australia**	
Post Office Address	Post Office Address **3199 Waverly Street**	City **Palo Alto**	State/Country **CA**	Zip Code **94306**
Full Name of Inventor 2	Last Name **BERNO**	First Name **Anthony**	Middle Name or Initial	
Residence & Citizenship	City **San Jose**	State/Foreign Country **California**	Country of Citizenship **Canada**	
Post Office Address	Post Office Address **570 South 12th street**	City **San Jose**	State/Country **CA**	Zip Code **95112**
Full Name of Inventor 3	Last Name **YANG**	First Name **Robert**	Middle Name or Initial	
Residence & Citizenship	City **San Ramon**	State/Foreign Country **California**	Country of Citizenship **United States**	
Post Office Address	Post Office Address **317 Winged Terrace**	City **San Ramon**	State/Country **CA**	Zip Code **94583**

I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Signature of Inventor 1	Signature of Inventor 2	Signature of Inventor 3
Mark Chee	Anthony Berno	Robert Yang
Date	Date	Date

h:\home\jol\work\16528\250-1\dpoa

Figure 1

A

5' ..TGAACTGTATCCGACAT..
3'   tgacat ggctgtag
          tgacatCggctgtag
          tgacatGggctgtag
          tgacatTggctgtag
3'   gacataAgctgtaga
           gacataCgctgtaga
           gacata gctgtaga
           gacataTgctgtaga

B

5' ..TGAACTGTACCCGACAT..
3'   tgacatAggctgtag
          tgacatCggctgtag
          tgacat ggctgtag
          tgacatTggctgtag
3'   gacataAgctgtaga
           gacataCgctgtaga
           gacata gctgtaga
           gacataTgctgtaga

C

5' TGAACTGTATCCGACAT

5' TGAACTGTACCCGACAT

16.493

Figure 2

Figure 3